stats

MDPI

Article

Affine Calculus for Constrained Minima of the Kullback–Leibler Divergence

Giovanni Pistone [1,2]

[1] De Castro Statistics, Collegio Carlo Alberto, 10122 Torino, Italy; giovanni.pistone@carloalberto.org
[2] Nuovo SEFIR, c/o Coworld, Centro Direzionale Milano Due, Palazzo Canova, 20054 Segrate, Italy

Abstract: The non-parametric version of Amari's dually affine Information Geometry provides a practical calculus to perform computations of interest in statistical machine learning. The method uses the notion of a statistical bundle, a mathematical structure that includes both probability densities and random variables to capture the spirit of Fisherian statistics. We focus on computations involving a constrained minimization of the Kullback–Leibler divergence. We show how to obtain neat and principled versions of known computations in applications such as mean-field approximation, adversarial generative models, and variational Bayes.

Keywords: information geometry; Kullback–Leibler divergence; statistical bundle; natural gradient

check for updates

Academic Editor: Wei Zhu

Received: 6 February 2025
Revised: 17 March 2025
Accepted: 19 March 2025
Published: 21 March 2025

Citation: Pistone, G. Affine Calculus for Constrained Minima of the Kullback–Leibler Divergence. *Stats* **2025**, *8*, 25. https://doi.org/10.3390/stats8020025

Copyright: © 2025 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction and Notations

Many modern Artificial Intelligence (AI) and machine learning (ML) algorithms are based on non-parametric statistical methods and optimization algorithms based on the minimization of a divergence measure between probability functions. In particular, one computes the gradient of a function defined on the probability simplex; then, the learning uses a gradient ascent technique. Such a basic approach is illustrated, for example, in the textbook [1] (Ch. 18).

In most papers, ordinary convex calculus tools on the open probability simplex provide the relevant derivatives and gradients. The relation between the analytic computations and their statistical meaning is not exposed. This paper focuses on the derivative and gradient computations by providing the geometric framework called Information Geometry (IG). This geometry differs from the usual convex analysis because its devices have a direct statistical meaning. For example, the velocity of a one-dimensional parametric curve $\theta \mapsto p(\theta)$ in the open probability simplex is defined to be the Fisher's score $\frac{d}{d\theta} \log p(\theta)$ instead of the ordinary derivative $\frac{d}{d\theta} p(\theta)$. Generally speaking, IG is a geometric interpretation of Fisherian inference ([1], Ch. 5).

Amari's Information Geometry (IG) [2–4] has been successfully applied to modern AI algorithms; see, for example, [5]. Here, we use the non-parametric version of IG of [6,7]. This version is *non-parametric* because the basic set of states is the open probability simplex, it is *affine*, as it satisfies a generalization of the classical Weyl's axioms [8]. Moreover, it is *dually affine* in the sense already defined in Amari contributions because the covariance bilinear operator appears as a duality pairing in the vector space of coordinates.

The specific applications we will consider as examples come from the literature in statistical ML, particularly those that involve the constrained minimization of the Kullback–Leibler divergence (KL-divergence). Indeed, our main result in Section 2 is a form of the

total gradient of the KL-divergence as expressed in the dually affine geometry. Namely, we consider symmetric divergences [9], generative adversarial networks [10], mixed entropy and transport optimization [11,12], and variational Bayes [13,14].

Non-parametric IG stands in general sample spaces and under various functional assumptions. One option, among many, is the use of Orlicz spaces [15]; see [6,16]. In this paper, we are not interested in discussing the functional setup. Still, we are interested in presenting the peculiar affine calculus of positive probability functions on a finite state space Ω in a geometric language compatible with the infinite-dimensional theory [17]. Such a calculus provides principled definitions of a curve's *velocity*, the scalar field's *gradient*, and the *gradient flow*.

1.1. Prerequisites

Below, we provide a schematic summary of the theory. For complete details, we refer to previous presentations in [7,18].

Let Ω be a finite sample space. We look at the open simplex as the *maximal exponential model* denoted as $\mathcal{E}(\Omega)$. In fact, we present every couple of positive probability functions on Ω, say p, q, in the form inspired by Statistical Physics [19]:

$$q = \mathrm{e}^{v-\Psi} \cdot p \,, \tag{1}$$

where p represents a ground state, q is a perturbation of the ground state, v is a random variable, Ψ is a normalizing constant, and $\psi = \kappa_p(v) = \log \mathbb{E}_p[\mathrm{e}^v]$ is the cumulant function.

The random variable v depends on p and q up to a constant. If we specify $\mathbb{E}_p[v] = 0$ in Equation (1), then a straightforward computation gives

$$\begin{aligned} v &= \log \frac{q}{p} - \mathbb{E}_p\left[\log \frac{q}{p}\right] \,, \quad q = \mathrm{e}^{v-\kappa_p(v)} \cdot p \,, \\ \kappa_p(v) &= \log \mathbb{E}_p[\mathrm{e}^v] = \mathbb{E}_p\left[\log \frac{p}{q}\right] = \mathrm{D}(p \,\|q) \,, \end{aligned}$$

where $\mathrm{D}(p \,\|q)$ is the KL-divergence. Regarding the entropy,

$$\mathrm{D}(p \,\|q) = \mathbb{E}_p\left[\log \frac{p}{q}\right] = \mathbb{E}_p[\log p] - \mathbb{E}_p[\log q] = -\,\mathcal{H}(p) + \mathcal{H}(p,q) \,.$$

If we specify $\mathbb{E}_q[v] = 0$ in Equation (1), an analogous computation gives

$$\begin{aligned} v &= \log \frac{q}{p} - \mathbb{E}_q\left[\log \frac{q}{p}\right] \,, \quad q = \mathrm{e}^{v-\kappa_p(v)} \cdot p \,, \\ \kappa_p(v) &= \log \mathbb{E}_p[\mathrm{e}^v] = -\,\mathbb{E}_q\left[\log \frac{p}{q}\right] = -\,\mathrm{D}(q \,\|p) \,. \end{aligned}$$

A vector bundle is a collection of vector spaces, and each vector space is called a *fiber* of the bundle. For example, the tangent bundle collects all tangent vectors at each point in differential geometry. In Fisher's statistics of the open probability simplex, one considers the vector space of all Fisher's scores of one-dimensional models through the probability function q. Inspired by this last example, we call the *statistical bundle* the vector bundle $S\mathcal{E}(\Omega)$ of all couples (q, v) of a positive probability function q and a q-centered random variable, $\mathbb{E}_q[v] = 0$,

$$S\mathcal{E}(\Omega) = \left\{(q, v) \,\middle|\, q \in \mathcal{E}(\Omega), \mathbb{E}_q[v] = 0\right\} \,.$$

Each fiber $S_q\mathcal{E}(\Omega)$ is a Euclidean space for the covariance inner product $\langle v, w\rangle_q = \mathbb{E}_q[vw]$.

The covariance inner product is both a Riemannian metric and a duality pairing. The metric interpretation leads to the Riemannian version of IG. The duality pairing interpretation leads to our dually affine IG. Because of that, we want to distinguish between the fibers $S_p\,\mathcal{E}(\Omega)$ and the dual fibers ${}^*S_p\,\mathcal{E}(\Omega)$. The first bundle is called *exponential bundle*, while the second bundle is called *mixture bundle*. We use the notation

$$ {}^*S_p\,\mathcal{E}(\Omega) \times S_p\,\mathcal{E}(\Omega) \ni (v,w) \mapsto \langle v,w\rangle_p\,, \quad p \in \mathcal{E}(\Omega)\,. $$

In our setup, all the vector spaces of random variables are finite-dimensional; hence, the fibers $S_p\,\mathcal{E}(\Omega)$ and ${}^*S_p\,\mathcal{E}(\Omega)$ are equal vector spaces. However, it is a useful distinction, as it will be apparent in the discussion of parallel transports below.

The definition of the statistical bundle aims to capture an essential mechanism of Fisher's approach to statistics ([1], Ch. 4). Suppose $t \mapsto q(t) \in \mathcal{E}(\Omega)$ is a one-dimensional statistical model. In that case, the Fisher's score is $t \mapsto \frac{d}{dt}\log q(t) = \dot{q}(t)$, and $t \mapsto (q(t), \dot{q}(t)) \in S\,\mathcal{E}(\Omega)$ is the *lift* of the curve to the statistical bundle.

Dually affine geometry follows from the definition of two parallel transports on the fibers and two affine charts. The parallel transports act between the fibers

$$ S_q\,\mathcal{E}(\Omega) \ni v \mapsto {}^{\mathrm{e}}\mathbb{U}^r_q v = v - \mathbb{E}_r[v] \in S_r\,\mathcal{E}(\Omega) \quad \textit{is the exponential transport,} \tag{2} $$

$$ {}^*S_q\,\mathcal{E}(\Omega) \ni w \mapsto {}^{\mathrm{m}}\mathbb{U}^r_q w = \frac{q}{r} v \in {}^*S_r\,\mathcal{E}(\Omega) \quad \textit{is the mixture transport.} \tag{3} $$

It is easy to check that the transports are duals of each other:

$$ \left\langle v, {}^{\mathrm{e}}\mathbb{U}^q_r w \right\rangle_q = \left\langle {}^{\mathrm{m}}\mathbb{U}^r_q v, w \right\rangle_r \quad \textit{is the transport's duality,} \tag{4} $$

$$ \langle v,w\rangle_q = \left\langle {}^{\mathrm{m}}\mathbb{U}^r_q v, {}^{\mathrm{e}}\mathbb{U}^r_q w \right\rangle_r \quad \textit{is the inner product push.} \tag{5} $$

The affine charts that define the two dual affine geometries by mapping the base set to a vector space of coordinates are

$$ q \mapsto s_p(q) = \log\frac{q}{p} - \mathbb{E}_p\left[\log\frac{q}{p}\right] \in S_p\,\mathcal{E}(\Omega) \quad \textit{is the exponential chart} \tag{6} $$

$$ q \mapsto \eta_p(q) = \frac{q}{p} - 1 \in {}^*S_p\,\mathcal{E}(\Omega) \quad \textit{is the mixture chart,} \tag{7} $$

and the geometries defined by the two atlases are affine because the parallelogram law holds in both cases:

$$ s_p(q) + {}^{\mathrm{e}}\mathbb{U}^p_q s_q(r) = s_p(r) $$
$$ \eta_p(q) + {}^{\mathrm{m}}\mathbb{U}^p_q \eta_q(r) = \eta_p(r) $$

The inverse of the exponential chart is a non-parametric exponential family ([1], Ch. 5), and the known mechanisms of the cumulant function provide a fundamental calculus tool [20]. If K_p is the restriction of κ_p to $S_p\,\mathcal{E}(\Omega)$ and $v = s_p(q)$ then $q = s_p^{-1}(v) = e_p(v) = \mathrm{e}^{v - K_p(v)} \cdot p$,

$$ K_p(v) = \log \mathbb{E}_p[\mathrm{e}^v] \quad \textit{cumulant function} \tag{8} $$

$$ \mathrm{D}\big(p \,\big\|\, e_p(v)\big) = K_p(v) \quad \textit{cumulant function express the KL- divergence,} \tag{9} $$

$$ dK_p(v)[h] = \mathbb{E}_{e_p(v)}[h] \quad \textit{is the derivative of } K_p \textit{ in the direction } h, \tag{10} $$

$$ d^2K(v)[h,k] = \mathrm{Cov}_{e_p(v)}(h,k) \quad \textit{is the second derivative of } K_p \textit{ in the directions } h,k. \tag{11} $$

Equations (10) and (11) are the non-parametric version of the well-known properties of the derivative of the cumulant function in exponential models; see ([1], § 5.5) and [20].

We can now show that Fisher's score is a velocity in the technical sense of a velocity computed in the moving frame of both charts. If $t \mapsto q(t) \in \mathcal{E}(\Omega)$ is a smooth curve, and $\Phi\colon\ \mathcal{E}(\Omega) \to \mathbb{R}$ is a smooth mapping,

$$\overset{\star}{q}(t) = \frac{d}{dt} s_p(q(t))\Big|_{p=q(t)} = \frac{d}{dt}\eta_p(q(t))\Big|_{p=q(t)} = \frac{d}{dt}\log q(t) \quad \textit{is the velocity,} \tag{12}$$

$$\frac{d}{dt}\Phi(q(t)) = \left\langle \operatorname{grad}\Phi(q(t)), \overset{\star}{q}(t)\right\rangle_{q(t)} \quad \textit{is the gradient.} \tag{13}$$

The squared norm of the velocity (12),

$$\left\langle \overset{\star}{q}(t), \overset{\star}{q}(t)\right\rangle_{q(t)} = \mathbb{E}_{q(t)}\left[\left(\frac{d}{dt}\log q(t)\right)^2\right] = \sum \frac{\dot{q}(t)^2}{q(t)}\,,$$

is the Fisher information that appeared first in the classical Cramer–Rao lower bound.

The gradient defined in Equation (13) is frequently called the *natural gradient* in the IG literature, following the use introduced in the case of parametric models by Amari [3]. In Riemannian geometry [17,21], the metric acts as a duality pairing, and the definition of the gradient is similar to Equation (13). The classic example of the computation of the gradient is the gradient of the expected value as a function of the probability function,

$$\frac{d}{dt}\,\mathbb{E}_{q(t)}[u] = \frac{d}{dt}\sum u\frac{\dot{q}(t)}{q(t)}\,q(t) = \left\langle u - \mathbb{E}_{q(t)}[u], \overset{\star}{q}(t)\right\rangle_{q(t)}\,,$$

so that $\operatorname{grad}\mathbb{E}_q[u] = u - \mathbb{E}_q[u]$.

The gradient of Φ gives the velocities of curves "orthogonal" to the surfaces of constant Φ-value, that is, the curves of steepest ascent. The solutions of the equation $\operatorname{grad}\Phi = 0$ are the stationary points of Φ, and an equation of the form

$$\overset{\star}{q}(t) = \epsilon(t)\operatorname{grad}\Phi(q(t))$$

is a *gradient flow equation*.

In conclusion, we review the derivation of a function f between two maximal exponential models using the mixture charts Equation (7). The expressions of f and its derivative df in the charts centered, respectively, at p_1 and p_2, are

$$\begin{array}{ccc} \mathcal{E}(\mu_1) & \xrightarrow{\ f\ } & \mathcal{E}(\mu_2) \\ {\scriptstyle n_{p_1}^{-1}}\uparrow & & \downarrow{\scriptstyle \eta_{p_2}} \\ S_{p_1}\mathcal{E}(\mu_1) & \xrightarrow[f_{p_1\cdot p_2}]{} & S_{p_2}\mathcal{E}(\mu_2) \end{array} \quad \text{and} \quad \begin{array}{ccc} S_{q_1}\mathcal{E}(\mu_1) & \xrightarrow{\ df(q_1)\ } & S_{f(q_1)}\mathcal{E}(\mu_2) \\ {\scriptstyle {}^{\mathrm{m}}\mathbb{U}_{p_1}^{q_1}}\uparrow & & \downarrow{\scriptstyle {}^{\mathrm{m}}\mathbb{U}_{f(p_1)}^{q_2}} \\ S_{p_1}\mathcal{E}(\mu_1) & \xrightarrow[df_{p_1\cdot p_2}(\eta_{p_1}(q_1))]{} & S_{p_2}\mathcal{E}(\mu_2) \end{array}$$

It follows that the computation of the derivative from its expression is

$$df(q)[\overset{\star}{q}] = {}^{\mathrm{m}}\mathbb{U}_{p_2}^{q} df_{p_1,p_2}(\eta_{p_1}(q))[{}^{\mathrm{m}}\mathbb{U}_{q}^{p_1}\overset{\star}{q}]\,. \tag{14}$$

1.2. Summary of Content

In the following sections, we give both new results and new versions of the known results. The aim is to show the interest of the non-parametric dually affine IG in computing the gradient flow of a constrained KL-divergence.

In Section 2, we show how to use the statistical bundle formalism to compute derivatives of functions defined on the open probability simplex and how to compute natural gradients and total natural gradients of the KL-divergence, the cross entropy, the entropy, and the Jensen–Shannon divergence.

In Section 3, we apply the general computations of the previous section to independence models and marginal conditional probabilities in a factorial product setting. The dually affine methodology methodically reproduces known computations and suggests neat variations of potential interest. In particular, Section 3.5 contains a fully worked example of the derivation of a gradient flow equation of interest in approximate Bayes computations.

2. Total Natural Gradient of the KL-Divergence

The KL-divergence ([5], Ch. 3) as a function of two variables is

$$D\colon\ \mathcal{E}(\Omega)\times\mathcal{E}(\Omega)\ni(q,r)\mapsto \mathrm{D}(q\,\|r)=\mathbb{E}_q\left[\log\frac{q}{r}\right]. \tag{15}$$

The computation of the total derivative is well-known in Information Theory. However, we provide proof in the affine setting, expressing the result in the affine charts.

In the exponential chart at p and in the mixture chart at p, the expressions of the probability functions q and r are, respectively,

$$q=e_p(v)=\mathrm{e}^{v-K_p(v)}\cdot p\,,\quad r=\eta_p^{-1}(w)=(1+w)\cdot p\,, \tag{16}$$

By plugging (16) into (15) and using Equation (10), one sees that the expressions of the partial KL-divergences are, respectively,

$$\begin{aligned}\mathrm{D}\big(e_p(v)\,\big\|r\big)=\mathbb{E}_{e_p(v)}\big[v-K_p(v)\big]-\mathbb{E}_{e_p(v)}\left[\log\frac{r}{p}\right]=\\ \mathbb{E}_{e_p(v)}\big[v-K_p(v)\big]-\mathbb{E}_{e_p(v)}\big[s_p(v)\big]-\mathrm{D}(p\,\|r)=\\ dK_p(v)[v]-K_p(v)-dK_p(v)[s_p(r)]-\mathrm{D}(p\,\|r)\,,\end{aligned} \tag{17}$$

and

$$\mathrm{D}\Big(q\,\Big\|\eta_p^{-1}(w)\Big)=\mathrm{D}(q\,\|p)-\mathbb{E}_q[\log(1+w)]\,. \tag{18}$$

Notice that the peculiar choice of the charts in the combination exponential for the first variable and mixture for the second variable is inessential in the finite state space case because any other choice will produce the same final result in the computation of the total natural gradient. However, it is consistent with the dual affine setting, in which two connections exist between one space and its dual. However, the expression of the KL-divergence using the exponential chart in both variables is interesting because, in such a case, the resulting expression is equal to the Bregman divergence of the cumulant function K_p,

$$\mathrm{D}\big(e_p(v)\,\big\|e_p(w)\big)=K_p(w)-K_p(v)-dK_p(v)[w-v]\,,$$

which, in turn, is the second remainder in the Taylor expansion. For example, one closed form is

$$\mathrm{D}\big(e_p(v)\,\big\|e_p(w)\big)=\int_0^1\int_0^1 d^2K(v+st(w-v))[w-v,w-v]\,sds\,dt\,.$$

If $q(s,t)=e_p(v+st(w-v))\propto q^{1-st}r^{st}$, then by Equation (11),

$$\mathrm{D}(q\,\|r)=\iint_0^1\mathrm{Var}_{q(s,t)}\left(\log\frac{r}{q}\right)s\,ds\,dt\,.$$

2.1. Total Natural Gradient of the KL-Divergence

We compute our gradients in the duality induced on each fiber by the covariance; hence, the *total natural gradient* of the KL-divergence has two components implicitly defined by

$$\frac{d}{dt}\,\mathrm{D}(q(t)\,\|r(t)) = \big\langle \overset{\star}{q}(t), \mathrm{grad}_1\,\mathrm{D}(q(t)\,\|r(t))\big\rangle_{q(t)} + \big\langle \mathrm{grad}_2\,\mathrm{D}(q(t)\,\|r(t)), \overset{\star}{r}(t)\big\rangle_{r(t)}\,, \quad (19)$$

where $\mathrm{grad}_1\,\mathrm{D}(q\,\|r)$ is a random variable in the fiber at Q, while $\mathrm{grad}_2\,\mathrm{D}(q\,\|r)$ is a random variable in the fiber at r. The adjective total refers to the fact that D is a function of two variables.

Proposition 1. *The total natural gradient of the KL-divergence is*

$$(q,r) \mapsto \mathrm{grad}\,\mathrm{D}(q\,\|r) = \big(-s_q(r), -\eta_r(q)\big) \in S\,\mathcal{E}(\Omega) \times {}^*S\,\mathcal{E}(\Omega)\,. \quad (20)$$

That is, more explicitly, for each smooth couple of curves $t \mapsto q(t)$ and $t \mapsto r(t)$, Equation (19) becomes

$$\frac{d}{dt}\,\mathrm{D}(q(t)\,\|r(t)) = -\Big\langle \overset{\star}{q}(t), s_{q(t)}(r(t))\Big\rangle_{q(t)} - \Big\langle \eta_{r(t)}(q(t)), \overset{\star}{r}(t)\Big\rangle_{r(t)}\,.$$

Proof. From Equation (11), the derivative at $v \in S_p\,\mathcal{E}(\Omega)$ of Equation (17) in the direction $h = {}^{\mathrm{e}}\mathbb{U}^p_q \overset{\star}{q}$ is

$$\begin{aligned} d^2K_p(v)[v,h] - dK_p(v)[h] + dK_p(v)[h] - d^2K_p(v)[s_p(r),h] = \\ \mathrm{Cov}_q\big(s_p(q) - s_p(r), h\big) = \mathrm{Cov}_q\Big(\log\frac{q}{r} - \mathbb{E}_p\Big[\log\frac{q}{r}\Big], h\Big) = \\ \mathbb{E}_q\Big[\Big(\log\frac{q}{r} - \mathbb{E}_q\Big[\log\frac{q}{r}\Big]\Big)\big(h - \mathbb{E}_q[h]\big)\Big] = \big\langle -s_q(r), \overset{\star}{q}\big\rangle_q\,. \end{aligned}$$

The derivative at $w \in {}^*S_p\,\mathcal{E}(\Omega)$ of Equation (18) in the direction $k = {}^{\mathrm{m}}\mathbb{U}^p_r \overset{\star}{r}$ is

$$-\,\mathbb{E}_q\Big[\frac{p}{r}k\Big] = -\,\mathbb{E}_r\Big[\frac{q}{r}\overset{\star}{r}\Big] = \big\langle -\eta_r(q), \overset{\star}{r}\big\rangle_r\,.$$

□

The gradient computation forms the corresponding gradient flow equation, whose discretization provides basic optimization algorithms. Here are two basic examples.

Given $r \in \mathcal{E}(\Omega)$, the solution of the gradient flow equation

$$\overset{\star}{q}(t) = -\,\mathrm{grad}_1\,\mathrm{D}(q(t)\,\|r) = s_q(r)\,, \quad q(0) = q_0\,,$$

is the exponential family

$$t \mapsto q(t) = \mathrm{e}^{\mathrm{e}^{-t}v_0 - K_r(\mathrm{e}^{-t}v_0)} \cdot r\,, \quad v_0 = s_r(q_0)\,. \quad (21)$$

In fact, the LHS of Equation (21) is given by

$$\begin{aligned} \log q(t) &= \mathrm{e}^{-t}v_0 - K_r(\mathrm{e}^{-t}v_0) + \log r\,, \\ \overset{\star}{q}(t) &= -\mathrm{e}^{-t}v_0 + dK_r(\mathrm{e}^{-t}v_0)[\mathrm{e}^{-t}v_0]\,, \end{aligned}$$

while the RHS is

$$s_{q(t)}(r) = -\mathrm{e}^{-t}v_0 + K_r(\mathrm{e}^{-t}v_0) - \mathbb{E}_{q(t)}\left[-\mathrm{e}^{-t}v_0 + K_r(\mathrm{e}^{-t}v_0)\right] = \\ -\mathrm{e}^{-t}v_0 + \mathbb{E}_{q(t)}\left[\mathrm{e}^{-t}v_0\right] .$$

The conclusion follows from Equation (10).

Given $q \in \mathcal{E}(\Omega)$, the solution of the gradient flow equation

$$\overset{\star}{r}(t) = -\operatorname{grad}_2 \mathrm{D}(q \,\|r(t)) = \eta_{r(t)}(q) \,, \quad r(0) = r_0 \,,$$

is the mixture family

$$t \mapsto r(t) = \mathrm{e}^{-t}r_0 + (1 - \mathrm{e}^{-t})q \,. \tag{22}$$

The LHS of Equation (22) is

$$\overset{\star}{r}(t) = \frac{\dot{r}(t)}{r(t)} = \frac{-\mathrm{e}^{-t}r_0 + \mathrm{e}^{-t}q}{\mathrm{e}^{-t}r_0 + (1 - \mathrm{e}^{-t})q} = \frac{q - r_0}{r_0 + (\mathrm{e}^{t} - 1)q} \,,$$

while the RHS is

$$\eta_{r(t)}(q) = \frac{q}{\mathrm{e}^{-t}r_0 + (1 - \mathrm{e}^{-t})q} - 1 = \frac{q - \mathrm{e}^{-t}r_0 + (1 - \mathrm{e}^{-t})q}{\mathrm{e}^{-t}r_0 + (1 - \mathrm{e}^{-t})q} = \frac{q - r_0}{r_0 + (\mathrm{e}^{t} - 1)q} \,.$$

Notice that in both cases, the t parameter appears in the solution in exponential form. Other forms of the temperature parameter will follow from a weighted form of the gradient flow equation.

2.2. Natural Gradient of the Entropy and Total Natural Gradient of the Cross Entropy

The KL-divergence equals the cross entropy minus the entropy,

$$\mathrm{D}(q \,\|r) = \mathbb{E}_q[-\log r] - \mathbb{E}_q[-\log q] = \mathcal{H}(q, r) - \mathcal{H}(q) \,.$$

In the exponential chart at p for the first variable, the cross entropy is

$$\mathcal{H}\left(s_p^{-1}(v), r\right) = \\ \mathbb{E}_{e_p(v)}[-\log r] = \mathbb{E}_{e_p(v)}\left[-\log r - \mathbb{E}_p[-\log r]\right] + \mathcal{H}(p, r) = \\ dK_p(v)[-\log r - \mathbb{E}_p[-\log r]] + \mathcal{H}(p, r) \,, \tag{23}$$

with derivative at v in the direction h

$$d^2K_p(v)[-\log r - \mathbb{E}_p[-\log r], h] = \mathrm{Cov}_q(-\log r, h) = \\ \mathbb{E}_q\left[(-\log r - \mathcal{H}(q, r))\left(h - \mathbb{E}_q[h]\right)\right] = \left\langle \overset{\star}{q}, -\log r - \mathcal{H}(q, r)\right\rangle_q \,.$$

In the mixture chart at p for the second variable

$$\mathcal{H}\left(q, \eta_p^{-1}(w)\right) = \mathbb{E}_q[-\log((1 + w) \cdot p)] = \mathbb{E}_q[-\log(1 + w)] + \mathcal{H}(q, p) \,, \tag{24}$$

with derivative at w in the direction k,

$$\mathbb{E}_q\left[-(1 + w)^{-1}k\right] = \mathbb{E}_r\left[-\frac{q}{r}\,{}^{\mathrm{m}}\mathbb{U}_p^r k\right] = \mathbb{E}_r\left[-\left(\frac{q}{r} - 1\right)\overset{\star}{r}\right] = \left\langle \eta_r(q), \overset{\star}{r}\right\rangle_r \,.$$

Proposition 2. *The total natural gradient of the cross entropy is*

$$\operatorname{grad}\mathcal{H}(q,r) = (-\log r - \mathcal{H}(q,r), \eta_r(q))$$

and the natural gradient of the entropy is

$$\operatorname{grad}\mathcal{H}(q) = -\log q - \mathcal{H}(q)\,. \tag{25}$$

Proof. The first statement follows from Equations (23) and (24). From the decomposition $\mathcal{H}(q) = \mathcal{H}(q,r) - \mathrm{D}(q\,\|r)$, we find the gradient of the entropy,

$$\operatorname{grad}\mathcal{H}(q) = -\log r - \mathcal{H}(q,r) + s_q(r) = -\log q - \mathcal{H}(q)\,.$$

□

2.3. Total Natural Gradient of the Jensen–Shannon Divergence

The *Jensen–Shannon* divergence [9] is

$$\mathrm{JS}(q,r) = \frac{1}{2}\,\mathrm{D}\left(q\,\middle\|\,\frac{1}{2}(q+r)\right) + \frac{1}{2}\,\mathrm{D}\left(r\,\middle\|\,\frac{1}{2}(q+r)\right) = \mathcal{H}\left(\frac{1}{2}(q+r)\right) - \frac{1}{2}\,\mathcal{H}(q) - \frac{1}{2}\,\mathcal{H}(r)\,. \tag{26}$$

It is the minimum value of the function

$$\phi\colon p \mapsto \frac{1}{2}(\mathrm{D}(q\,\|p) + \mathrm{D}(r\,\|p))\,.$$

In fact,

$$\operatorname{grad}\phi(p) = -\frac{1}{2}\left(\eta_p(q) + \eta_p(r)\right) = \frac{1}{2}\left(-\frac{q}{p} + 1 - \frac{r}{p} + 1\right) = -\frac{\frac{1}{2}(q+r)}{p} + 1$$

which vanishes for $p = \frac{1}{2}(q+r)$.

Let us compute the derivative of $f\colon q \mapsto \frac{1}{2}(q+r)$. The mixture expression of f at p according to Equation (7) is the affine function

$$f_p(v) = \eta_p \circ f \circ \eta_p^{-1}(v) = \frac{\frac{1}{2}(1+v)\cdot p + r}{p} - 1 = \frac{1}{2}v + \frac{1}{2}\eta_p(r)\,,$$

so that the derivative in the direction h is $df_p(v)[h] = h/2$.

The push-back, according to the mixture transport Equation (3), is

$$df(q)[\dot{q}] = {}^{\mathrm{m}}\mathbb{U}_p^{\frac{1}{2}(q+r)}\,\frac{1}{2}\,{}^{\mathrm{m}}\mathbb{U}_q^p\,\dot{q} = \frac{p}{\frac{1}{2}(q+r)}\,\frac{1}{2}\,\frac{q}{p}\,\dot{q} = \frac{1}{2}\,\frac{q}{\frac{1}{2}(q+r)}\,\dot{q} = \frac{1}{2}\,{}^{\mathrm{m}}\mathbb{U}_q^{\frac{1}{2}(q+r)}\,\dot{q}\,. \tag{27}$$

We now compute the gradient of $q \mapsto \mathrm{JS}(q,r)$ of Equation (26), using the total natural gradient of the KH-divergence of Proposition 1, the derivative Equation (27), and the duality of parallel transports Equation (4):

$$\begin{aligned}
&\operatorname{grad}(q \mapsto \mathrm{JS}(q,r)) = \\
&\quad \frac{1}{2}\left(-s_q\left(\frac{1}{2}(q+r)\right) - \frac{1}{2}\,{}^{\mathrm{e}}\mathbb{U}^q_{\frac{1}{2}(q+r)}\eta_{\frac{1}{2}(q+r)}(q)\right) + \frac{1}{2}\left(-\frac{1}{2}\,{}^{\mathrm{e}}\mathbb{U}^q_{\frac{1}{2}(q+r)}\eta_{\frac{1}{2}(q+r)}(r)\right) = \\
&\qquad \frac{1}{2}\left(-s_q\left(\frac{1}{2}(q+r)\right) - \frac{1}{2}\,{}^{\mathrm{e}}\mathbb{U}^q_{\frac{1}{2}(q+r)}\left(\eta_{\frac{1}{2}(q+r)}(q) + \eta_{\frac{1}{2}(q+r)}(r)\right)\right) = \\
&\quad \frac{1}{2}\left(-s_q\left(\frac{1}{2}(q+r)\right) - \frac{1}{2}\,{}^{\mathrm{e}}\mathbb{U}^q_{\frac{1}{2}(q+r)}\left(\frac{q}{\frac{1}{2}(q+r)} - 1 + \frac{r}{\frac{1}{2}(q+r)} - 1\right)\right) = \\
&\qquad\qquad -\frac{1}{2}s_q\left(\frac{1}{2}(q+r)\right).
\end{aligned}$$

It is also instructive to use the expression of the Jensen–Shannon divergence in terms of entropies. From Equation (25),

$$\begin{aligned}
&\operatorname{grad}(q \mapsto \mathrm{JS}(q,r)) = \\
&\quad \frac{1}{2}\,{}^{\mathrm{e}}\mathbb{U}^q_{\frac{1}{2}(q+r)}\left(-\log\frac{1}{2}(q+r) - \mathcal{H}\left(\frac{1}{2}(q+r)\right)\right) - \frac{1}{2}(-\log q - \mathcal{H}(q)) = \\
&\qquad -\frac{1}{2}\left(\log\frac{1}{2}(q+r) - \mathbb{E}_q\left[\log\frac{1}{2}(q+r)\right] + \log q - \mathbb{E}_q[\log q]\right) = \\
&\qquad\qquad -\frac{1}{2}s_q\left(\frac{1}{2}(q+r)\right).
\end{aligned}$$

3. Product Sample Space

This section uses $\Omega = \Omega_1 \times \Omega_2$ as a factorial sample space. For each $r \in \mathcal{E}(\Omega)$, the margins are $r_1 \in \mathcal{E}(\Omega_1)$ and $r_2 \in \mathcal{E}(\Omega_2)$. In the *mean-field assumption*, the model equals the tensor product of the margins,

$$\bar{r} = r_1 \otimes r_2 \in \mathcal{E}(\Omega_1) \otimes \mathcal{E}(\Omega_2) \subset \mathcal{E}(\Omega)\,.$$

The velocities are, respectively,

$$\overset{\star}{r}(x,y;t) = \frac{d}{dt}\log r(x,y;t)\,, \tag{28}$$

$$\overset{\star}{\bar{r}}(x,y;t) = \frac{d}{dt}\log \bar{r}(x,y;t) = \overset{\star}{r}_1(x;t) + \overset{\star}{r}_2(y;t)\,. \tag{29}$$

Below, we will discuss the optimality of a mean-field approximation.

3.1. Product Sample Space: Marginalization

The (first) marginalization is

$$\Pi_1\colon\ \mathcal{E}(\Omega_1 \times \Omega_2) \ni r \mapsto r_1 \in \mathcal{E}(\Omega_1)\,, \quad r_1(x) = \sum_b r(x,b)\,. \tag{30}$$

We will compute the bundle derivative of Equation (30) following the scheme of Equation (14).

Proposition 3. *The derivative* $d\Pi_1$ *of the marginalization Equation* (30) *is*

$$d\Pi_1\colon\ {}^{*S}\mathcal{E}(\Omega) \ni (r,\overset{\star}{r}) \mapsto \left(r_1, \mathbb{E}_r\left(\overset{\star}{r}\middle|\Pi_1\right)\right) \in {}^{*S}\mathcal{E}(\Omega_1)\,.$$

Proof. In the mixture chart centered at $p_1 \otimes p_2$ and p_1, respectively, the expression of the marginalization is

$$\eta_{p_1} \circ \Pi_1 \circ \eta^{-1}_{p_1 \otimes p_2}(v) = \frac{\Pi_1 \circ \eta^{-1}_{p_1 \otimes p_2}(v)}{p_1} - 1 = \frac{\sum_b (1 + v(\cdot, b)) \cdot p_1 p_2(b)}{p_1} - 1 = \sum_b v(\cdot, b) p_2(b) \,. \quad (31)$$

Note that the expression in Equation (31) is linear. Hence, the derivative at v in the direction h is $x \mapsto \sum_b h(\cdot, b) p_2(b)$ with $h = {}^{\mathrm{m}}\mathbb{U}_r^{p_1 \otimes p_2} \overset{\star}{r}$ so that the bundle derivative is

$$d\Pi_1(r)[\overset{\star}{r}] = {}^{\mathrm{m}}\mathbb{U}_{p_1}^{r_1} \sum_b \frac{r(\cdot, b)}{p_1 p_2(b)} \overset{\star}{q}(\cdot, b) p_2(b) = \frac{p_1}{r_1} \sum_b \frac{r(\cdot, b)}{p_1 p_2(b)} \overset{\star}{q}(\cdot, b) p_2(b) = \sum_b \overset{\star}{q}(\cdot, b) r_{1|2}(\cdot | b) = \mathbb{E}_r\left(\overset{\star}{r} \middle| \Pi_1\right) \,.$$

□

There is an interesting relation between conditional expectation and mixture transport. The conditional expectation commutes with the mixture transports,

$${}^{\mathrm{m}}\mathbb{U}_{r_1}^{q_1} \mathbb{E}_r\left({}^{\mathrm{m}}\mathbb{U}_q^r v \middle| \Pi_1\right) = \mathbb{E}_q(v | \Pi_1) \,, \quad v \in {}^*S_q \, \mathcal{E}(\Omega) \,.$$

It is a way to express Bayes' theorem for conditional expectations. For all ϕ,

$$\mathbb{E}_q\left[{}^{\mathrm{m}}\mathbb{U}_{r_1}^{q_1} \mathbb{E}_r\left({}^{\mathrm{m}}\mathbb{U}_q^r v \middle| \Pi_1\right) \phi(\Pi_1)\right] = \mathbb{E}_{q_1}\left[\frac{r_1}{q_1} \mathbb{E}_r\left(\frac{q}{r} v \middle| \Pi_1\right) \phi(\Pi_1)\right] =$$
$$\mathbb{E}_{r_1}\left[\mathbb{E}_r\left(\frac{q}{r} v \middle| \Pi_1\right) \phi(\Pi_1)\right] = \mathbb{E}_r\left[\mathbb{E}_r\left(\frac{q}{r} v \middle| \Pi_1\right) \phi(\Pi_1)\right] = \mathbb{E}_r\left[\frac{q}{r} v \Phi(\Pi_1)\right] = \mathbb{E}_q[v \phi(\Pi_1)] \,.$$

3.2. Product Sample Space: Mean-Field Approximation

The derivative of the joint marginalization

$$\Pi \colon \mathcal{E}(\Omega_1 \times \Omega_2) \ni r \mapsto r_1 \Pi_1(r) \otimes \Pi_2(r) \in \mathcal{E}(\Omega_1 \times \Omega_2) \,,$$

follows from the derivative of the marginalization in Equation (29).

Proposition 4. *The derivative* $d\Pi$ *of the joint marginalization in* Section 3.2 *is*

$$d\Pi \colon (r, \overset{\star}{r}) \mapsto \left(r_1 \otimes r_2, \mathbb{E}_r\left(\overset{\star}{r} \middle| \Pi_1\right) + \mathbb{E}_r\left(\overset{\star}{r} \middle| \Pi_2\right)\right) \,. \quad (32)$$

Proof. Compose the partial derivatives with the mapping $(\overset{\star}{r}_1, \overset{\star}{r}_2) \mapsto \overset{\star}{r}_1 \otimes \overset{\star}{r}_2$ □

The decomposition of the velocity *velocity* r according to Equation (32) provides a better decomposition than $\overset{\star}{r}_1 + \overset{\star}{r}_2$ of Equations (28) and (29) and provides a definition of the *mean-field approximation*. In the language of ANOVA decomposition of statistical interactions, the derivative part in Equation (32) is the sum of the simple effects of the velocity,

$$\overset{\star}{q} = \overset{\star}{q}_1 + \overset{\star}{q}_2 + \overset{\star}{q}_{12} \,,$$

where $q_i = \mathbb{E}_q\left(\overset{\star}{q} \middle| \Pi_i\right)$, $i = 1, 2$, and the last term is the interaction, the q-orthogonal residual. See [22] for a discussion of the ANOVA decomposition in the context of the statistical bundle.

The equation for the total natural gradient of the KL-divergence and the computation of the derivative above provide the natural gradients of the divergence between the joint probability function and the mean-field approximation. In information theory [23], the KL-divergence in Equation (34) is called mutual information.

Proposition 5. *The natural gradients of the divergences of a joint distribution r and its mean-field approximation $\Pi(r)$ are*

$$\operatorname{grad} \mathrm{D}(\Pi(r) \,\|r) = \mathbb{E}_{r_1 \otimes r_2}(s_r(r_1 \otimes r_2)|\Pi_1) + \mathbb{E}_{r_1 \otimes r_2}(s_r(r_1 \otimes r_2)|\Pi_2) - \eta_r(r_1 \otimes r_2)\,. \quad (33)$$

$$\operatorname{grad} \mathrm{D}(r \,\|\Pi(r)) = -s_r(r_1 \otimes r_2) + \mathbb{E}_r(\eta_r(r_1 \otimes r_2)|\Pi_1) + \mathbb{E}_r(\eta_r(r_1 \otimes r_2)|\Pi_2)\,. \quad (34)$$

The conditional terms in Equation (33) depend on the mean-field model; hence, we could express them as a disintegration of r. For example,

$$\mathbb{E}_{r_1 \otimes r_2}(s_r(r_1 \otimes r_2)|\Pi_1 = x) = \sum_y r_2(y) \log \frac{r_1(x) r_2(y)}{r(x,y)} - \sum_{xy} r_1(x) r_2(y) \log \frac{r_1(x) r_2(y)}{r(x,y)} = \mathcal{H}\Big(r_2, r_{2|1}(\cdot|x)\Big) - \sum_x r_1(x)\, \mathcal{H}\Big(r_2, r_{2|1}(\cdot|x)\Big)\,,$$

where the last term is the conditional entropy $\mathcal{H}(\Pi_2|\Pi_1)$ under r.

Proof of Equation (33). We find the natural gradient of $r \mapsto \mathrm{D}(\Pi(r) \,\|r)$ by computing with Equations (20) and (32) the variation along a smooth curve $t \mapsto r(t) \in \mathcal{E}(\Omega_1 \times \Omega_2)$ such that $r(0) = r$ and $\dot{r}(0) = \dot{r}$. It holds that

$$\frac{d}{dt} \mathrm{D}(\Pi(r(t)) \,\|r(t))\Big|_{t-0} = -\Big\langle s_{\Pi(r)}(r), d\Pi(r)[\dot{r}] \Big\rangle_{\Pi(r)} - \big\langle \eta_r(\Pi(r)), \dot{r} \big\rangle_r = - \big\langle s_{r_1 \otimes r_2}(r), \mathbb{E}_r\big(\dot{r}\big|\Pi_1\big) + \mathbb{E}_r\big(\dot{r}\big|\Pi_2\big) \big\rangle_{r_1 \otimes r_2} - \big\langle \eta_r(r_1 \otimes r_2), \dot{r} \big\rangle_r$$

We want to present the first term of the RHS as an inner product at r applied to $\dot{r}$. Let us push the inner product from $r_1 \otimes r_2$ to r with Equation (5). It holds that

$$\begin{aligned} &- \big\langle s_{r_1 \otimes r_2}(r), (\mathbb{E}_r\big(\dot{r}\big|\Pi_1\big) + \mathbb{E}_r\big(\dot{r}\big|\Pi_2\big)) \big\rangle_{r_1 \otimes r_2} = \\ &\quad - \big\langle {}^{\mathrm{e}}\mathbb{U}^r_{r_1 \otimes r_2} s_{r_1 \otimes r_2}(r), {}^{\mathrm{m}}\mathbb{U}^r_{r_1 \otimes r_2}(\mathbb{E}_r\big(\dot{r}\big|\Pi_1\big) + \mathbb{E}_r\big(\dot{r}\big|\Pi_2\big)) \big\rangle_r = \\ &\quad \Big\langle s_r(r_1 \otimes r_2), \frac{r_1 \otimes r_2}{r} (\mathbb{E}_r\big(\dot{r}\big|\Pi_1\big) + \mathbb{E}_r\big(\dot{r}\big|\Pi_2\big)) \Big\rangle_r = \\ &\quad \mathbb{E}_r\Big[s_r(r_1 \otimes r_2) \frac{r_1 \otimes r_2}{r} (\mathbb{E}_r\big(\dot{r}\big|\Pi_1\big) + \mathbb{E}_r\big(\dot{r}\big|\Pi_2\big)) \Big] = \\ &\mathbb{E}_r\Big[\Big(\mathbb{E}_r\Big(s_r(r_1 \otimes r_2) \frac{r_1 \otimes r_2}{r} \Big| \Pi_1 \Big) + \mathbb{E}_r\Big(s_r(r_1 \otimes r_2) \frac{r_1 \otimes r_2}{r} \Big| \Pi_2 \Big) \Big) \dot{r} \Big] = \\ &\quad \big\langle \mathbb{E}_{r_1 \otimes r_2}(s_r(r_1 \otimes r_2)|\Pi_1) + \mathbb{E}_{r_1 \otimes r_2}(s_r(r_1 \otimes r_2)|\Pi_2), \dot{r} \big\rangle_r\,. \end{aligned}$$

The last equality follows from

$$\mathbb{E}_r\Big(s_r(r_1 \otimes r_2) \frac{r_1 \otimes r_2}{r} \Big| \Pi_i \Big) = \mathbb{E}_{r_1 \otimes r_2}(s_r(r_1 \otimes r_2)|\Pi_i)\,, \quad i = 1, 2\,.$$

□

Proof of Equation (34).

$$\frac{d}{dt}\,\mathrm{D}(r(t)\,\|\Pi(r(t)))\Big|_{t=0} = -\left\langle s_r(r_1\otimes r_2),\overset{\star}{r}\right\rangle_r - \left\langle \eta_{r_1\otimes r_2}(r), d\Pi(r)[\overset{\star}{r}]\right\rangle_{r_1\otimes r_2} = \\ -\left\langle s_r(r_1\otimes r_2),\overset{\star}{r}\right\rangle_r - \left\langle \eta_{r_1\otimes r_2}(r), \mathbb{E}_r\left(\overset{\star}{r}\middle|\Pi_1\right)+\mathbb{E}_r\left(\overset{\star}{r}\middle|\Pi_2\right)\right\rangle_{r_1\otimes r_2}$$

and compute the second term as

$$-\left\langle \eta_{r_1\otimes r_2}(r), \mathbb{E}_r\left(\overset{\star}{r}\middle|\Pi_1\right)+\mathbb{E}_r\left(\overset{\star}{r}\middle|\Pi_2\right)\right\rangle_{r_1\otimes r_2} = \\ -\left\langle {}^{\mathrm{m}}\mathbb{U}^r_{r_1\otimes r_2}\eta_{r_1\otimes r_2}(r), {}^{\mathrm{e}}\mathbb{U}^r_{r_1\otimes r_2}\left(\mathbb{E}_r\left(\overset{\star}{r}\middle|\Pi_1\right)+\mathbb{E}_r\left(\overset{\star}{r}\middle|\Pi_2\right)\right)\right\rangle_r = \\ \left\langle \eta_r(r_1\otimes r_2), \mathbb{E}_r\left(\overset{\star}{r}\middle|\Pi_1\right)+\mathbb{E}_r\left(\overset{\star}{r}\middle|\Pi_2\right)\right\rangle_r = \\ \mathbb{E}_r\left[\eta_r(r_1\otimes r_2)\left(\mathbb{E}_r\left(\overset{\star}{r}\middle|\Pi_1\right)+\mathbb{E}_r\left(\overset{\star}{r}\middle|\Pi_2\right)\right)\right] = \\ \mathbb{E}_r\left[\left(\mathbb{E}_r(\eta_r(r_1\otimes r_2)|\Pi_1)+\mathbb{E}_r(\eta_r(r_1\otimes r_2)|\Pi_2)\right)\overset{\star}{r}\right] = \\ \left\langle \mathbb{E}_r(\eta_r(r_1\otimes r_2)|\Pi_1)+\mathbb{E}_r(\eta_r(r_1\otimes r_2)|\Pi_2),\overset{\star}{r}\right\rangle_r .$$

Equation (34) follows. □

3.3. Product Sample Space: Kantorovich and Scrödinger

If Π denotes the joint marginalization, the set of *transport plans* with margins q_1 and q_2 is

$$\Gamma(q_1,q_2) = \Phi^{-1}(q_1\otimes q_2) = \{q\,|\,\Pi(q) = q_1\otimes q_2\}\,.$$

Here, we deal with a classical topic with considerable literature. We mention only the monograph of ref. [12] and, from the Information Geometry perspective, ref. [11,22].

Let us consider first the Kantorovich problem. Given the cost function (i.e., potential function)

$$U\colon \Omega\to\mathbb{R}\,,$$

and a curve $t\mapsto q(t)\in\gamma(q_1,q_2)$, we want to minimize the cost

$$S(t) = \mathbb{E}_{q(t)}[U]\,.$$

As $\mathbb{E}_{q(t)}[\phi(\Pi_i)] = \mathbb{E}_{q_i}[\phi]$ for all ϕ,

$$0 = \frac{d}{dt}\left\langle U-\mathbb{E}_{q(t)}[U],\overset{\star}{q}(t)\right\rangle_{q(t)},$$

so that $\mathbb{E}_{q(t)}\left(\overset{\star}{q}(t)\middle|\Pi_i\right) = 0$. The velocity of a curve in the transport plans is an interaction. Now, the derivative of the cost is

$$\frac{d}{dt}S(q(t)) = \left\langle U-\mathbb{E}_{q(t)}[U],\overset{\star}{q}(t)\right\rangle_{q(t)}. \tag{35}$$

From the interaction property of $\overset{\star}{q}(t)$, it follows that if the ANOVA decomposition

$$U = \mathbb{E}_q[U] + (u_1(\Pi_1;q)+u_2(\Pi_2;q)) + u_{12}(\Pi_1,\Pi_2;q)$$

holds, then

$$\frac{d}{dt}S(q(t)) = \left\langle u_{12}(\Pi_1,\Pi_2;q(t)),\overset{\star}{q}(t)\right\rangle_{q(t)}.$$

The Scrödinger problem is similar. Given the cost function (i.e., potential function)

$$U\colon \Omega\to\mathbb{R}\,,$$

consider the exponential perturbation of the mean-field probability function

$$\exp\left(-\frac{U}{\epsilon}-\phi(\epsilon)\right)\cdot q_1\otimes q_2\,. \tag{36}$$

The parameter $\epsilon>0$ is called *temperature*, and the normalizing constant is

$$\psi(\epsilon)=\log \mathbb{E}_{q_1\otimes q_2}\left[\mathrm{e}^{-U/\epsilon}\right].$$

The KL-divergence of q relative to the perturbed probability function of Equation (36) is

$$S_\epsilon(q)=\mathrm{D}\left(q\,\middle\|\,\mathrm{e}^{-U/\epsilon-\phi(\epsilon)}\cdot(q_1\otimes q_2)\right)=$$
$$\mathbb{E}_q\left[\log\frac{q}{\log q_1\otimes q_2}\right]+\epsilon^{-1}\,\mathbb{E}_q[U]+\psi(\epsilon)=$$
$$\epsilon^{-1}(S(q)+\epsilon\,\mathrm{D}(q\,\|q_1\otimes q_2)+\epsilon\phi(\epsilon))\,.$$

The gradient of $q\mapsto S_\epsilon(q)$ is, from Equations (35) and (34),

$$\operatorname{grad} S_\epsilon(q)=$$
$$\epsilon^{-1}\left(U-\mathbb{E}_q[U]-s_q(q_1\otimes q_2)+\mathbb{E}_q\left(\eta_q(q_1\otimes q_2)\middle|\Pi_1\right)+\mathbb{E}_q\left(\eta_q(q_1\otimes q_2)\middle|\Pi_2\right)\right).$$

Only the interaction part is relevant in the constrained problem $q\in\Gamma(q_1,q_2)$, and the interaction kills the two conditional expectations, which leaves

$$\left(U-\mathbb{E}_q[U]\right))_{12;q}-\left(s_q(q_1\otimes q_2)\right)_{12;q}.$$

We refer to [22] for a method to compute the interaction part of a random variable.

3.4. Product Sample Space: Conditional Probability Function

When the sample space is a product, $\Omega=\Omega_1\times\Omega_2$, we can represent each probability function in the maximal exponential model via conditioning on one margin,

$$\mathcal{E}(\Omega_1\times\Omega_2)=\left\{q=q_{1|2}\cdot q_2\,\middle|\,q_{1|2}(\cdot|y)\in\mathcal{E}(\Omega_1),y\in\Omega_2,q_2\in\mathcal{E}(\Omega_2)\right\} \tag{37}$$
$$=\left\{q=q_{2|1}\cdot q_1\,\middle|\,q_{2|1}(\cdot|x)\in\mathcal{E}(\Omega_2),x\in\Omega_1,q_1\in\mathcal{E}(\Omega_1)\right\}. \tag{38}$$

The two representations are

$$\mathcal{E}(\Omega_1)^{\otimes\Omega_2}\times\mathcal{E}(\Omega_2)\leftrightarrow\mathcal{E}(\Omega_1\times\Omega_2)\leftrightarrow\mathcal{E}(\Omega_1)\times\mathcal{E}(\Omega_2)^{\otimes\Omega_1}\,. \tag{39}$$

Following the approach of [10], ([5], Ch. 11), we look at the transition mapping

$$\Omega_2\ni y\mapsto q_{1|2}(\cdot|y)\in\mathcal{E}(\Omega_1)$$

as a family of probability functions representing alternative probability models. The other transition mapping

$$\Omega_1\ni x\mapsto q_{2|1}(\cdot|x)\in\mathcal{E}(\Omega_2)$$

is the *discriminator*, that is, $q_{2|1}(y|x)$ is the probability that the sample x comes from $q_{1|2}(\cdot|y)$.

The right-to-left second mapping in Equation (39),

$$B\colon\ \mathcal{E}(\Omega_1)\times\mathcal{E}(\Omega_2)^{\otimes\Omega_1}\ni\left(q_1,q_{2|1}(\cdot|x),x\in\Omega_1\right)\mapsto q=q_{2|1}\cdot q_1\in\mathcal{E}(\Omega_1\times\Omega_2)\,, \tag{40}$$

maps the vector of the 1-margin and the set of alternative probability functions to the joint probability function. The kinematics of Equation (40), that is, the computation of velocities, is

$$\dot{q}(x,y;t) = \dot{q}_1(x;t) + \dot{q}_{2|1}(y|x;t)\,,$$

Hence, the total derivative of B is

$$dB(q_1, q_{2|1}(\cdot|x), x \in \Omega_1)[\dot{q}_1, \dot{q}_{2|1}(\cdot|x), x \in \Omega_1] \colon (x,y) \mapsto \dot{q}_1(x) + \dot{q}_{2|1}(y|x)\,.$$

The transposed total derivative is defined by

$$\Big\langle v, dB(q_1, q_{2|1})[\dot{q}_1, \dot{q}_{2|1}] \Big\rangle_{B(q_1,q_{2|1})} = \Big\langle dB(q_1, q_{2|1})^*[v], (\dot{q}_1, \dot{q}_{2|1}) \Big\rangle_{(q_1,q_{2|1})}\,,$$

that is,

$$\begin{aligned}
&\sum_{x,y} v(x,y)\Big(\dot{q}_1(x) + \dot{q}_{2|1}(y|x)\Big) q_1(x) q_{2|1}(y|x) = \\
&\qquad \sum_{x,y} v(x,y)\dot{q}_1(x) q_1(x) q_{2|1}(y|x) + \sum_{x,y} v(x,y)\dot{q}_{2|1}(y|x) q_1(x) q_{2|1}(y|x) = \\
&\qquad \sum_x \left(\sum_y v(x,y) q_{2|1}(y|x) \right) \dot{q}_1(x) q_1(x) + \sum_x \sum_y q_1(x) v(x,y) \dot{q}_{2|1}(y|x) q_{2|1}(y|x) = \\
&\qquad\qquad \big\langle \mathbb{E}_q(v|\Pi_1), \dot{q}_1 \big\rangle_{q_1} + \sum_x q_1(x) \sum_y v(x,y) \dot{q}_{2|1}(y|x) q_{2|1}(y|x) = \\
&\qquad\qquad \big\langle \mathbb{E}_q(v|\Pi_1), \dot{q}_1 \big\rangle_{q_1} + \sum_x \Big\langle q_1(x)\Big(v(x,\cdot) - \mathbb{E}_{q_{2|1}(\cdot|x)}[v(x,\cdot)]\Big), \dot{q}_{2|1}(y|x) \Big\rangle_{q_{2|1}(\cdot|x)}\,.
\end{aligned}$$

In conclusion, the transposed total derivative is

$$dB(q_1, q_{2|1})^*[v] \colon (x,y) \mapsto \Big(\mathbb{E}_q(v|\Pi_1), q_1(x)\Big(v(x,\cdot) - \mathbb{E}_{q_{2|1}(\cdot|x)}[v(x,\cdot)]\Big), x \in \Omega_1\Big)\,.$$

It is interesting to derive dB in the mixture atlas. The mixture expression of B with respect to $(p_1, p_2^{\otimes \Omega_1})$ and $p_1 \otimes p_2$ is

$$\begin{aligned}
B_{p_1,p_2} \colon (v_1, v_{2|1}(\cdot|x), x \in \Omega_1) \mapsto \\
(x,y) \mapsto (1 + v_1(x)) p_1(x) (1 + v_{2|1}(y|x)) p_2(y) \mapsto \\
(1 + v_1)(1 + v_{2|1}) - 1
\end{aligned}$$

and the total derivative in the directions $h_1, h_{2|1}$ is

$$\begin{aligned}
dB_{p_1,p_2}(v_1, v_{2|1}(\cdot|z), z \in \Omega_1)[h_1, h_{2|1}(\cdot|z), z \in \Omega_1] = \\
(1 + v_{2|1}) h_1 + \sum_z (1 + v_1(z)) h_{2|1}(\cdot|z)\,.
\end{aligned}$$

The push-back of the total derivative expression to the statistical bundles uses the equations

$$1 + v_1 = q_1/p_1, 1 + v_{2|1} = q_{2|1}/p_2, h_1 = {}^{\mathrm{m}}\mathbb{U}^{p_1}_{q_1} \dot{q}_1, h_{2|1}(\cdot|z) = {}^{\mathrm{m}}\mathbb{U}^{p_2}_{q_{2|1}(\cdot|z)} \dot{q}_{2|1}(\cdot|z)$$

to obtain

$$dB(q_1, q_{2|1}(\cdot|z), z \in \Omega_1)[\overset{\star}{q}_1, \overset{\star}{q}_{2|1}(\cdot|z), z \in \Omega_1] = {}^{\mathrm{m}}\mathbb{U}^{q_{2|1}\cdot q_1}_{p_1\otimes p_2}\left(\frac{q_{2|1}}{p_2}\,{}^{\mathrm{m}}\mathbb{U}^{p_1}_{q_1}\overset{\star}{q}_1 + \sum_z \frac{q_1(z)}{p_1(z)}\,{}^{\mathrm{m}}\mathbb{U}^{p_2}_{q_{2|1}(\cdot|z)}\overset{\star}{q}_{2|1}(\cdot|z)\right) = \frac{p_1\otimes p_2}{q_{2|1}\cdot q_1}\left(\frac{q_{2|1}}{p_2}\frac{q_1}{p_1}\overset{\star}{q}_1 + \sum_z \frac{q_1(z)}{p_1(z)}\frac{q_{2|1}(\cdot|z)}{p_2}\overset{\star}{q}_{2|1}(\cdot|z)\right)$$

In our affine language, we repeat computations in [10]. We especially derive the natural gradient of a composite function by the equation

$$\operatorname{grad}\Phi\circ B = dB^*\operatorname{grad}\Phi(B)\,.$$

For a given target $p \in \mathcal{E}(\Omega_1\times\Omega_2)$, we express the KL-divergence as a function of B in Equation (40),

$$K_q\colon (r_1, r_{2|1}(\cdot|x), x\in\Omega_1) \mapsto \mathrm{D}\Big(p\,\Big\|\,B(q_1, q_{2|1}(\cdot|x), x\in\Omega_1)\Big) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{q_1(x)q_{2|1}(y|x)}\,.$$

We have

$$\operatorname{grad} K_p(B(q_1, q_{2|1})) = -\eta_{q_1\cdot q_{2|1}}(p) = 1 - \frac{p}{q_1\cdot q_{2|1}}\,.$$

The first component of $\operatorname{grad} K_p\circ B$ is

$$[dB(q_1, q_{2|1})^*\operatorname{grad} K_p(B(q_1, q_{2|1}))]_1\colon x\mapsto \sum_z\left(1-\frac{p(x,z)}{q_1(x)q_{2|1}(z|x)}\right)q_{2|1}(z|x) = \sum_z\frac{q_1(x)q_{2|1}(z|x)-p(x,z)}{q_1(x)} = 1-\frac{p_1(x)}{q_1(x)}\,, \quad (41)$$

so that $[\operatorname{grad} K_p\circ B(q_1, q_{2|1})]_1 = -\eta_{q_1}(p_1)$. The x-component is

$$[dB(q_1, q_{2|1})^*\operatorname{grad} K_p(B(q_1, q_{2|1}))]_x\colon y\mapsto q_1(x)\left(1-\frac{p(x,y)}{q_1(x)q_{2|1}(y|x)}\right) - q_1(x)\sum_y\left(1-\frac{p(x,y)}{q_1(x)q_{2|1}(y|x)}\right)q_{2|1}(y|x) = -\frac{p(x,y)}{q_{2|1}(y|x)} + p_1(x)\,, \quad (42)$$

so that $[dB(q_1, q_{2|1})^*\operatorname{grad} K_p(B(q_1, q_{2|1}))]_x = -p_1(x)\eta_{q_{2|1}(\cdot|x)}(p_{2|1}(\cdot|x))$.

We now assume a target probability function $g\in\mathcal{E}(\Omega_1)$ and consider the probability function on the product sample space where all the model probability functions equal g, and the discriminator is uniform, say $p = \frac{1}{m}g$, $m = \#\Omega_2$.

In this case, Equation (41) becomes

$$\operatorname{grad} K_p\circ B(q_1, q_{2|1})]_1 = -\eta_{q_1}(g)$$

and Equation (42) becomes

$$\operatorname{grad} K_p\circ B(q_1, q_{2|1})]_x = -q_1(x)\left(1-\frac{g(x)/m}{q_1(x)q_{2|1}(y|x)}\right)$$

3.5. Variational Bayes

We revisit and develop some computations of ([13], § 2.2). We keep the same notation as above so that Bayes' formula is

$$q_{2|1}(y|x) = \frac{q_{1|2}(x|y)q_2(y)}{q_1(x)} = \frac{q_{1|2}(x|y)q_2(y)}{\sum_y q_{1|2}(x|y)q_2(y)} ,$$

where x is a sample value and y is a latent variable value.

For a fixed $x \in \Omega_1$, we look for a r in some model $\mathcal{M} \subset \mathcal{E}(\Omega_2)$ in order to approximate the conditional $q_{2|1}(\cdot|x)$. If $q \mapsto \mathcal{L}(r, x)$ satisfies

$$\log q_1(x) = \mathrm{D}\Big(r \,\Big\|q_{2|1}(\cdot|x)\Big) + \mathcal{L}(r, x) ,$$

then

$$\begin{aligned}\mathcal{L}(r,x) = \log q_1(x) - \sum_y r(y) \log \frac{r(y)}{q_{2|1}(y|x)} = \\ \sum_y r(y)\left(\log q_1(x) - \log \frac{r(y)}{q_{2|1}(y|x)}\right) = \sum_y r(y) \log \frac{q_{2|1}(y|x)q_1(x)}{r(y)} = \\ \sum_y r(y) \log \frac{q_{12}(x,y)}{r(y)} = \sum_y r(y) \log \frac{q_2(y)}{r(y)} + \sum_y r(y) \log \frac{q_{12}(x,y)}{q_2(y)} = \\ - \mathrm{D}(r \,\|q_2) + \sum_y r(y) \log q_{1|2}(x|y) .\end{aligned}$$

The so-called *variational lower bound* follows from $\mathrm{D}\Big(r \,\Big\|q_{2|1}(\cdot|x)\Big) \geq 0$,

$$\begin{aligned}\log q_1(x) = \mathrm{D}\Big(r \,\Big\|q_{2|1}(\cdot|x)\Big) - \mathrm{D}(r \,\|q_2) + \mathbb{E}_r\Big[\log q_{1|2}(x|\cdot)\Big] \geq \\ - \mathrm{D}(r \,\|q_2) + \mathbb{E}_r\Big[\log q_{1|2}(x|\cdot)\Big] = \mathcal{L}(r; x) .\end{aligned}$$

for all $r \in \mathcal{M}$. The bound is *exact* because $r(y) = q_{2|1}(y|x)$ for all y if, and only if, $\mathrm{D}\Big(r \,\Big\|q_{2|1}(\cdot|x)\Big) = 0$ The lower-bound variation along a curve $t \mapsto r(t) \in \mathcal{M}$ is

$$\begin{aligned}\frac{d}{dt}\mathcal{L}(r(t); x) = \\ \Big\langle \dot{r}(t), s_{r(t)}(q_2)\Big\rangle_{r(t)} + \Big\langle \dot{r}(t), \log q_{1|2}(x|\cdot) - \mathbb{E}_{r(t)}\Big[\log q_{1|2}(x|\cdot)\Big]\Big\rangle_{r(t)} = \\ \left\langle \dot{r}(t), \log \frac{q_2}{r(t)} + \log q_{1|2}(x|\cdot) - \mathbb{E}_{r(t)}\left[\log \frac{q_2}{r} + \log q_{1|2}(x|\cdot)\right]\right\rangle_{r(t)} = \\ \left\langle \dot{r}(t), \log \frac{q_{12}(x,\cdot)}{r(t)} - \mathbb{E}_{r(t)}\left[\log \frac{q_{12}(x,\cdot)}{r(t)}\right]\right\rangle_{r(t)} .\end{aligned} \quad (43)$$

If the model $\mathcal{M}$ is an exponential tilting of the margin q_2,

$$r = e_{q_2}(\theta \cdot u) = \mathrm{e}^{\theta \cdot u - \psi(\theta)} \cdot q_2 ,$$

where $\theta \in \Theta \subset \mathbb{R}^d$ is a vector parameter, and u is the vector of sufficient statistics of the exponential family with $\mathbb{E}_{q_2}[u] = 0$, then the velocity in Equation (43) becomes

$$\dot{r}(t) = \dot{\theta}(t) \cdot (u - \nabla\psi(\theta(t)) = \dot{\theta}(t) \cdot (u - \mathbb{E}_{r(t)}[u]) = \dot{\theta}(t) \cdot {}^{\mathrm{e}}\mathbb{U}_{q_2}^{r(t)} u ,$$

and the gradient in Equation (43) becomes

$$\log \frac{q_{12}(x,\cdot)}{r(t)} - \mathbb{E}_{r(t)}\left[\log \frac{q_{12}(x,\cdot)}{r(t)}\right] =$$
$$\log \frac{q_{12}(x,\cdot)}{q_2} - (\theta(t)\cdot u - \psi(u)) - \mathbb{E}_{r(t)}\left[\log \frac{q_{12}(x,\cdot)}{q_2} - (\theta(t)\cdot u - \psi(u))\right] =$$
$$\log q_{1|2}(x|\cdot) - \theta(t)\cdot u - \mathbb{E}_{r(t)}\left[\log q_{1|2}(x|\cdot) - \theta(t)\cdot u\right] =$$
$${}^{\mathrm{e}}\mathbb{U}_{q_2}^{r(t)}\left(\log q_{1|2}(x|\cdot) - \mathbb{E}_{q_2}\left[\log q_{1|2}(x|\cdot)\right] - \theta(t)\cdot u\right).$$

Using repeatedly Equation (11), we find the derivative of the lower bound in Equation (43),

$$\frac{d}{dt}\mathcal{L}(r(t);x) =$$
$$\left\langle \dot\theta(t)\cdot {}^{\mathrm{e}}\mathbb{U}_{q_2}^{r(t)} u, {}^{\mathrm{e}}\mathbb{U}_{q_2}^{r(t)}\left(\log q_{1|2}(x|\cdot) - \mathbb{E}_{q_2}\left[\log q_{1|2}(x|\cdot)\right] - \theta(t)\cdot u\right)\right\rangle_{r(t)} =$$
$$\sum_{i=1}^{d}\dot\theta_i(t)\left(\mathrm{Cov}_{r(t)}\left(u_i, \log q_{1|2}(x|\cdot)\right) - \sum_{j=1}^{d}\theta_j(t)\,\mathrm{Cov}_{r(t)}\left(u_i,u_j\right)\right) =$$
$$\dot\theta(t)\cdot\left(\mathrm{Cov}_{r(t)}\left(u, \log q_{1|2}(x|\cdot)\right) - \mathrm{Hess}\,\psi(\theta(t))\theta(t)\right).$$

In conclusion, the gradient flow equation for the maximization of the lower bound under the model $\mathcal{M}$ is

$$\dot\theta = -\,\mathrm{Hess}\,\psi(\theta(t))\theta(t) + \mathrm{Cov}_{e_{q_2}(\theta(t)\cdot u)}\left(u, \log q_{1|2}(x|\cdot)\right). \tag{44}$$

As a sanity check, assume that the model is exact for the given x,

$$q_{2|1}(y|x) = \mathrm{e}^{\bar\theta\cdot u(y) - \psi(\bar\theta)}\cdot q_2(y).$$

Hence,

$$\log q_{1|2}(x|y) = \log \mathrm{e}^{\bar\theta\cdot u(y)-\psi(\bar\theta)} q_2(y)\frac{q_1(x)}{q_2(y)} = \bar\theta\cdot u = \psi(\bar\theta) + \log q_1(x),$$

where the last two terms do not depend on y so that

$$\mathrm{Cov}_{e_{q_2}(\theta\cdot u}\left(u, \bar\theta\cdot u\right) = \mathrm{Hess}\,\psi(\theta)\bar\theta$$

and Equation (44) becomes

$$\dot\theta = -\,\mathrm{Hess}\,\psi(\theta(t))(\theta(t) - \bar\theta).$$

The solution of the gradient flow Equation (44) requires the ability to compute the covariance for the current model distribution. We do not discuss the numerical and simulation issues related to the implementation here.

4. Discussion

In this paper, we have shown how the dually affine formalism for the open probability simplex provides a system of affine charts in which the statistical notion of Fisher's score becomes the moving-chart affine velocity, and the natural gradient becomes the gradient. The construction applies to standard computations of interest for statistical machine learning. In particular, we have discussed the neat form of the total gradient of the KL-divergence

and its applications in a factorial sample space, such as mean-field approximation and Bayes computations.

This approach is helpful because the unique features of the Fisherian approach to statistics, such as Fisher's score, maximum likelihood, and Fisher's information, are formalized as an affine calculus so that all the statistical tools are available in this more extensive theory. Moreover, this setting potentially unifies the formalisms of Statistics, Optimal Transport, and Statistical Physics, examples being the affine modeling of Optimal Transport [24] and the second-order methods of optimization [25].

We have not considered the implementation of the formal gradient flow equation as a practical learning algorithm. Such further development is currently outside the scope of this piece of research. It would require the numerical analysis of the continuous equation and the search for sampling versions of the expectation operators. We hope this note will prompt further research. On the abstract side, topics worth studying seem to be the cases of continuous sample space as in [16], or Gaussian models as in [26].

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study is purely methodological. No data were produced or used.

Acknowledgments: The Author acknowledges the partial support of de Castro Statistics, Collegio Carlo Alberto. He is a member of the non-profit organization Nuovo SEFIR and the UMI interest group AI&ML&MAT.

Conflicts of Interest: The Author declares no conflicts of interest.

References

1. Efron, B.; Hastie, T. *Computer Age Statistical Inference*; Institute of Mathematical Statistics (IMS) Monographs; Algorithms, evidence, and data science; Cambridge University Press: New York, NY, USA, 2016; Volume 5, pp. xix+475.
2. Amari, S. Geometry of Semiparametric Models and Applications. Invited Papers Meeting IP64 Likelihood and Geometry. Organizer Preben F. Blaesild. In Proceedings of the 51st Session of the International Statistical Institute, Istanbul, Turkey, 18–26 August 1997.
3. Amari, S.I. Natural gradient works efficiently in learning. *Neural Comput.* **1998**, *10*, 251–276. [CrossRef]
4. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; Translated from the 1993 Japanese original by Daishi Harada; American Mathematical Society: Providence, RI, USA, 2000; pp. x+206.
5. Amari, S.I. *Information Geometry and Its Applications*; Applied Mathematical Sciences; Springer: Tokyo, Japan, 2016; Volume 194, pp. xiii+374.
6. Pistone, G.; Sempi, C. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Statist.* **1995**, *23*, 1543–1561. [CrossRef]
7. Chirco, G.; Pistone, G. Dually affine Information Geometry modeled on a Banach space. *arXiv* **2022**, arXiv:2204.00917. [CrossRef]
8. Weyl, H. *Space- Time- Matter*; Translation of the 1921 RAUM ZEIT MATERIE; Dover: New York, NY, USA, 1952.
9. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
10. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *arXiv* **2014**, arXiv:1406.2661. [CrossRef]
11. Amari, S.i.; Karakida, R.; Oizumi, M. Information geometry connecting Wasserstein distance and Kullback-Leibler divergence via the entropy-relaxed transportation problem. *Inf. Geom.* **2018**, *1*, 13–37. [CrossRef]
12. Peyré, G.; Cuturi, M. Computational Optimal Transport. *Found. Trends Mach. Learn.* **2019**, *11*, 355–607. [CrossRef]
13. Kingma, D.P.; Welling, M. Auto-encoding variational Bayes. *arXiv* **2013**, arXiv:1312.6114.
14. Khan, M.E.; Rue, H.V. The Bayesian learning rule. *J. Mach. Learn. Res.* **2023**, *24*, 46.
15. Musielak, J. *Orlicz Spaces and Modular Spaces*; Lecture Notes in Mathematics; Springer: Berlin/Heidelberg, Germany, 1983; Volume 1034, pp. iii+222.
16. Pistone, G. Information geometry of the Gaussian space. In *Information Geometry and Its Applications*; Springer: Cham, Switzerland, 2018; Volume 252, pp. 119–155.

17. Lang, S. *Differential and Riemannian Manifolds*, 3rd ed.; Graduate Texts in Mathematics; Springer: Berlin/Heidelberg, Germany, 1995; Volume 160, pp. xiv+364.
18. Pistone, G. Information Geometry of the Probability Simplex: A Short Course. *Nonlinear Phenom. Complex Syst.* **2020**, *23*, 221–242. [CrossRef]
19. Landau, L.D.; Lifshits, E.M. *Course of Theoretical Physics*, 3rd ed.; Statistical Physics; Butterworth-Heinemann: Oxford, UK, 1980; Volume V.
20. Brown, L.D. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*; Number 9 in IMS Lecture Notes; Monograph Series; Institute of Mathematical Statistics: Ann Arbor, MI, USA, 1986; pp. x+283.
21. do Carmo, M.P. *Riemannian Geometry*; Mathematics: Theory & Applications; Translated from the Second Portuguese Edition by Francis Flaherty; Birkhäuser Boston Inc.: Berlin, Germany, 1992; pp. xiv+300.
22. Pistone, G. Statistical Bundle of the Transport Model. In *Geometric Science of Information*; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 752–759. [CrossRef]
23. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2006; pp. xxiv+748.
24. Ay, N. Information geometry of the Otto metric. *Inf. Geom.* **2024**, 1–24. [CrossRef]
25. Chirco, G.; Malagò, L.; Pistone, G. Lagrangian and Hamiltonian dynamics for probabilities on the statistical bundle. *Int. J. Geom. Methods Mod. Phys.* **2022**, *19*, 2250214. [CrossRef]
26. Malagò, L.; Montrucchio, L.; Pistone, G. Wasserstein Riemannian geometry of Gaussian densities. *Inf. Geom.* **2018**, *1*, 137–179. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.